DARM: Distribution-Aware Reward Modeling by Alleviating Biases from Low Preference-Context Dependency Data

Anonymous ACL submission

Abstract

Reward models (RMs) are the surrogate objectives in reinforcement learning from human feedback (RLHF), and their scores directly steer policy optimization. We show that standard RM training is vulnerable in data subsets where response quality depends only weakly on the context: such instances encourage the RM to ignore the context, leading to context neglect and degraded accuracy. To address this failure mode, we propose **D**istribution-**A**ware **R**eward **M**odeling (**DARM**), which augments the RM objective with a conditional mutual information regularizer that maximizes context and the predicted reward conditioned on the response. By explicitly preserving the sensitivity of reward signals to the prompting context, DARM reduces over-reliance on response-only features and improves robustness to contextual variation. Extensive experiments across in-distribution and out-of-distribution settings show that DARM trained RMs deliver more accurate and consistent scoring than strong baselines. We further evaluate its downstream impact in RLHF, where DARM produces better aligned policies. We also demonstrate the necessity of each DARM design component and the impact of key parameters on performance through ablation experiments.

1 Introduction

Reinforcement learning from human feedback (RLHF) is the key post-training paradigm for aligning large language models (LLMs) with human preferences (Ouyang et al., 2022; Bai et al., 2022a). In RLHF, a reward model (RM) serves as the learned surrogate objective: given a context (prompt) and two candidate responses, it predicts which response is preferred and supplies the reward signal for policy optimization. The training data for a RM typically consists of a context paired with two responses, one accepted and one rejected, and the objective of RM is to maximize the margin between their assigned rewards (Gao et al., 2023; Achiam et al., 2023). Under this objective, a well-trained RM would evaluate a candidate response with respect to both its conditioning context and the response itself (Wang et al., 2024; Bansal et al., 2024).

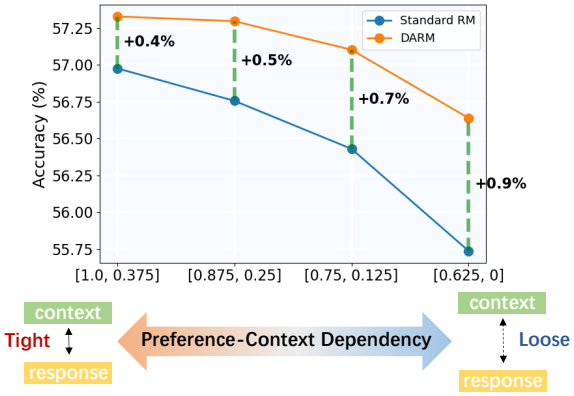


Figure 1: Test dataset accuracy for Standard RM and DARM trained on subsets stratified by preference-context dependency score. Dependency decreases from left to right, the x-axis shows quantile bins of each subsets scores, and the rightmost bin corresponds to the lowest preferencecontext dependency. DARM consistently yields larger gains in low-dependency regimes. Green lines denote the absolute accuracy lift of DARM over Standard RM.

However, real world preference datasets contain instances whose labels can be determined from the response alone. For example, when the response contains overtly offensive or unsafe content that annotators would reject regardless of the context. While such examples are valuable for aligning models with human values, they inadvertently encourage a response-only shortcut: the RM learns to underweight the context when scoring responses. This bias degrades accuracy and robustness precisely in situations where contextual information is decisive for judging response quality. To examine how the dependency between the response-

preference and the context in training data affects RM performance, we train otherwise identical models on subsets with varying degrees of preference-context dependency and compare their accuracy in the test dataset. Specifically, we employ a state-of-the-art open-source RM as a gold scorer and measure how sensitive its scores are to perturbations in the context. Treating this sensitivity as a preference-context dependency (PCD) score[1], we rank all training instances from low to high and form equal-sized subsets spanning different quantiles of the ranking. We then train separate RMs on each subset with matched sample counts. As shown in the blue curve in Figure 1, test accuracy degrades steadily as PCD of the training subset decreases, indicating that low PCD induces a context-neglect bias and undermines RM generalization. In contrast, the proposed DARM markedly mitigates this bias: as illustrated by the green line in Figure 1, the lower the subsets PCD, the larger DARMs accuracy gains over the Standard RM.

We next detail the DARM's training objective that yields these gains. Specifically, DARM augments the RM objective with a conditional mutual information (CMI) regularizer that strengthens the dependence between the reward signal and the context given the response, thereby encouraging the model to consult the context rather than rely on response-only shortcut. We estimate the CMI term with an InfoNCE-style contrastive objective (Oord et al., 2018): conditioned on the given response, the RM learns to identify the true context among conditionally matched fake variants. Negative contexts are constructed by randomly masking portions of the context while preserving the preference relation for the paired responses. Because the proposed method explicitly accounts for heterogeneous sample types within the data distribution, we refer to our approach as Distribution-Aware Reward Modeling (DARM).

We evaluate our approach on a suite of public out-of-distribution (OOD) benchmarks, measuring RM accuracy against strong baselines. We then assess downstream alignment by applying RLHF and comparing policies on dialogue and summarization tasks. Across settings, DARM consistently improves OOD RM accuracy and, in turn, yields better-aligned policies. We further conduct ablation studies to quantify the contribution of each DARM design component to overall performance. The main contributions of our paper are as follows:

- We identify a dataset-level source of context neglect in RM training: instances with weak preference-context dependency that bias RMs toward response-only shortcut.
- We introduce a distribution-aware reward modeling method DARM that augments the RM objective by encouraging a higher conditional mutual information term, thereby explicitly preserving preference-context dependency conditioned on responses.
- On public ID and OOD benchmarks, DARM consistently outperforms strong RM baselines and improves RLHF outcomes, producing better-aligned policies for dialogue and summarization tasks.

2 Related Work

2.1 Reinforcement Learning from Human Feedback and Reward Model

Reinforcement learning from human feedback (RLHF) has become the prevailing paradigm for aligning large language models (LLMs) with human preferences (Bai et al., 2022a; Ouyang et al., 2022; Zheng et al., 2023a). Beyond general alignment, RLHF has delivered consistent improvements across summarization, dialogue, and translation, while encouraging helpful, honest, and harmless behavior (Stiennon et al., 2020; Ziegler et al., 2019; Bahdanau et al., 2017; Thoppilan et al., 2022; Ouyang et al., 2022). The standard RLHF pipeline first trains a reward model (RM) on human preference pairs and then optimizes the policy to maximize the learned reward. The RM serves as the learned objective that transmits human preferences to the policy. Its biases and failure modesfor example, brittleness under distribution shiftdirectly shape downstream behavior (Bai et al., 2022b). Consequently, improving RM robustness and generalization is pivotal for RLHF effectiveness (Ramé et al., 2024; Lee et al., 2023).

Recent work seeks to improve RMs by steering them toward task-relevant signals and away from spurious cues. From a causal perspective, prior studies encourage decisions to rely on salient, causally pertinent information rather than nuisance factors (Liu et al., 2024a; Wang et al., 2025). From an information-theoretic perspective, Miao

[1]Details of the PCD computation appear at the beginning of the Appendix. We defer them here to preserve the flow of the main text.

et al. (2024) maximize the mutual information between intermediate representations and labels to suppress irrelevant interference. Complementary analyses reveal that RMs often under-attend to the prompt (context) when scoring responsesespecially on mathematical tasksthereby harming effectiveness and robustness under distribution shift (Bansal et al., 2024). To counter this, Dou et al. (2025) directly increase attention on context tokens to bolster robustness.

In this work, we identify a distinct, data-driven source of context neglect: training sets containing many samples with weak preference-context dependency bias RMs toward response-only shortcuts. We therefore aim to enhance fidelity to human preferences by mitigating this data induced bias, restoring dependence on the context when evaluating responses.

2.2 Contrastive Learning

Contrastive learning frames representation or policy learning as separating positives from negatives, typically via temperature-scaled softmax objectives (Oord et al., 2018). Seminal methods such as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) established strong empirical foundations for noise-contrastive estimation and instance discrimination, and continue to inform modern alignment losses. On the textual side, sentence-level contrastive methods (e.g., SimCSE/SBERT) highlight the importance of view construction for language, where dropout/back-translation stand in for image augmentations (Gao et al., 2021). A parallel thread studies the quality of negatives. Hard-negative mining and cross batch memory improve discrimination but exacerbate false negatives (semantically similar positives)(Wang et al., 2020). Debiased contrastive learning and importance reweighting mitigate this issue by down-weighting suspected collisions(Chuang et al., 2020; Kalantidis et al., 2020). These observations directly motivate our use of context masked negatives: they are semantically close to the original response yet differ precisely along the context dimension we wish the RM to attend to.

Recent work has applied contrastive objectives to alignment: Hejna et al. (2023) fine-tunes policies with a contrastive objective to obtain an RL-free alignment framework, and Chen et al. (2024) incorporates noise-contrastive estimation into DPO to leverage multi-candidate preference data. In this paper, we employ contrastive learning as an estimator of the lower bound of conditional mutual information(Oord et al., 2018; Song and Ermon, 2019; Mukherjee et al., 2020).

3 Method

In this section, we first formalize notation, then detail our method and provide an intuitive gradient-based explanation for its effectiveness.

3.1 Distribution-Aware Reward Modeling

Let $\mathcal{D}$ be a distribution over tuples (c, r^+, r^-), where c is the context, and r^+/r^- are the chosen/rejected responses. We introduce a hypothetical "golden" reward model $R^\star$, and define the context-sensitivity:

$$\Delta(c, r^+, r^-) \triangleq \left|R^\star(c) - R^\star(c, r^+)\right| + \left|R^\star(c) - R^\star(c, r^-)\right|. \tag{1}$$

Given a threshold $\tau > 0$, we partition the support of $\mathcal{D}$ into

$$\mathcal{D}_{\text{ctx}} \triangleq \left\{(c, r^+, r^-),\ \Delta(c, r^+, r^-) > \tau\right\},$$
$$\mathcal{D}_{\text{no-ctx}} \triangleq \left\{(c, r^+, r^-),\ \Delta(c, r^+, r^-) \leq \tau\right\}.$$

Intuitively, instances in $\mathcal{D}_{\text{no-ctx}}$ are "context-agnostic": the pairwise scores $R^\star(c, r^\pm)$ stay close to the context-only prior $R^\star(c)$, so the preference between r^+ and r^- is largely insensitive to c; in contrast, instances in $\mathcal{D}_{\text{ctx}}$ are "context-sensitive". We view $\mathcal{D}$ as a mixture with π as the ratio of $\mathcal{D}_{\text{ctx}}$:

$$\mathcal{D} = \pi \cdot \mathcal{D}_{\text{ctx}} + (1 - \pi) \cdot \mathcal{D}_{\text{no-ctx}}. \tag{2}$$

When π is small (i.e., $\mathcal{D}_{\text{no-ctx}}$ dominates), the RM tends to ignore the context when scoring, which degrades its performance on $\mathcal{D}_{\text{ctx}}$ and undermines robustness.

Our objective is to counter the bias introduced by $\mathcal{D}_{\text{no-ctx}}$ samples, ensuring that the RM properly conditions on the context (prompt) when evaluating a response. To this end, we augment the RMs training objective with a negative conditional mutual-information (CMI) regularizer:

$$-I\big(c; R_\theta(c, r^+) - R_\theta(c, r^-) \mid r^+, r^-\big), \tag{3}$$

to explicitly preserve label-context dependence given the response. Intuitively, forcing the CMI term quantity up eliminates response-only shortcuts and makes the reward margin depend on context whenever the human judgment does.

Let $\Delta_\theta(c) := R_\theta(c, r^+) - R_\theta(c, r^-)$, we estimate and maximize a contrastive lower bound to

$I\big(c; \Delta_\theta(c) \mid r^+, r^-\big)$ with the InfoNCE objective:

$$\mathcal{L}_{\text{NCE}}(\theta) = -\,\mathbb{E}\left[\log \frac{\exp\big(R_\theta(c, r^+)/\tau\big)}{\sum_{j=1}^{K} \exp\big(R_\theta(c_j, r^+)/\tau\big)}\right] + \mathbb{E}\left[\log \frac{\exp\big(R_\theta(c, r^-)/\tau\big)}{\sum_{j=1}^{K} \exp\big(R_\theta(c_j, r^-)/\tau\big)}\right]. \tag{4}$$

where θ denotes the RM parameters and c^* is the ground-truth context paired with the response pair (r^+, r^-). $C = \{c^*\} \cup \{c_1, \ldots, c_{K-1}\}$ is the candidate set formed by keeping (r^+, r^-) fixed and generating conditionally matched negatives. We combine the InfoNCE term with the Bradley Terry model (Bradley and Terry, 1952) preference loss using a single coefficient λ, and $\tau > 0$ is the temperature used to control the softmax sharpness. The overall DARM objective is:

$$\mathcal{L}_{\text{DARM}}(\theta) = \mathcal{L}_{\text{BT}}(\theta) \;+\; \lambda\, \mathcal{L}_{\text{NCE}}(\theta). \tag{5}$$

Equation 4 defines the contrastive logits via response-level pointwise rewards. A natural alternative is a pairwise CMI objective in Equation 6 that uses reward difference $R_\theta(c, r^+) - R_\theta(c, r^-)$ between the preferred and rejected responses as the contrastive logit:

$$\mathcal{L}_{\text{NCE}}(\theta) = -\mathbb{E}_{(c^*, r^+, r^-)}\left[\log \frac{\exp\big(\Delta_\theta(c^*)/\tau\big)}{\sum_{j=1}^{K} \exp\big(\Delta_\theta(c_j)/\tau\big)}\right]. \tag{6}$$

The pairwise margin formulation directly aligned with the standard pairwise preference objective used in RM training, avoids calibration issues associated with absolute reward scales, and integrates cleanly with the BT loss through a shared reward head. However, we observe that a pairwise formulation can degrade OOD accuracy. We present a head-to-head comparison between the pairwise and pointwise CMI estimation in Section 4.4.

Given the centrality of negative samples in contrastive learning, we explore two ways to generate negative samples:

- **Context-swap.** Replace $c^\star$ with a context c sampled from a different training instance, while keeping (r^+, r^-) fixed.
- **Span mask (1-of-N).** Partition $c^\star$ into N contiguous, equal-length segments. Uniformly sample one span and replace it with a mask placeholder, preserving special tokens and leaving the overall sequence length essentially unchanged.

Intuitively, the first introduces strong global mismatches that quickly suppress response-only shortcuts; the latter yields hard, distribution-faithful perturbations that directly increase sensitivity to localized evidence.

3.2 Why it works

Let $\mathcal{C} = \{c^\star, c_1, \ldots, c_{K-1}\}$ and define

$$p_j = \text{softmax}\big(\Delta_\theta(c_j)/\tau\big), \qquad \sum_{c \in \mathcal{C}} p_c = 1. \tag{7}$$

Then the partial derivative of Eq. (6) with respect to the margin of any candidate $c_j \in \mathcal{C}$ is

$$\frac{\partial \mathcal{L}_{\text{NCE}}}{\partial \Delta_\theta(c_j)} = \frac{1}{\tau}\Big(p_j - \mathbb{I}[c_j = c^\star]\Big). \tag{8}$$

Consequently, for any parameter block ϕ,

$$\nabla_\phi \mathcal{L}_{\text{NCE}} = \frac{1}{\tau} \sum_{c \in \mathcal{C}} \Big(p_c - \mathbb{I}[c = c^\star]\Big)\, \nabla_\phi \Delta_\theta(c). \tag{9}$$

Equation (9) is a centered, softmax-weighted average of margin gradients across the K candidates. In particular, the coefficients sum to zero, i.e., $\sum_j^K (p_j - \mathbb{I}[c = c^\star]) = 0$.

From a parameter decomposition perspective, we write the reward model as

$$R_\theta(c, r) = g_{\mathbf{w}_r}(r) \;+\; h_{\mathbf{w}_c}(c, r), \tag{10}$$

where $g_{\mathbf{w}_r}$ is a response-only pathway and $h_{\mathbf{w}_c}$ captures preference-context dependency.

Let $\mathcal{D}_{\text{no-ctx}}$ denote instances where preference is weakly dependent on the context (the margin is nearly context-invariant), and $\mathcal{D}_{\text{ctx}}$ denote instances where preference does hinge on context.

With the decomposition (10),

$$\Delta_\theta(c) = \underbrace{g_{\mathbf{w}_r}(r^+) - g_{\mathbf{w}_r}(r^-)}_{\text{independent of } c} + \underbrace{h_{\mathbf{w}_c}(c, r^+) - h_{\mathbf{w}_c}(c, r^-)}_{\text{depends on } c}. \tag{11}$$

Hence $\nabla_{\mathbf{w}_r} \Delta_\theta(c)$ is a constant vector w.r.t. c. Plugging into Equation (9),

$$\nabla_{\mathbf{w}_r} \mathcal{L}_{\text{NCE}} = \frac{1}{\tau}\left(\sum_{c \in \mathcal{C}} p_c - \sum_{c \in \mathcal{C}} \mathbb{I}[c = c^\star]\right) \nabla_{\mathbf{w}_r} \Delta_\theta(c) = \frac{1}{\tau}(1 - 1)\nabla_{\mathbf{w}_r} \Delta_\theta(c) = \mathbf{0}. \tag{12}$$

Thus, the loss update produces zero gradient on the response-only pathway ($\mathbf{w}_r$): it does not inflate response-only shortcuts, and all learning pressure

Model	WebGPT	SHP	RB	RMB-Helpful	RMB-Harmless	Average Scores	Diff
Standard RM	58.51%	52.19%	73.30%	63.63%	51.70%	59.87%	0.00%
WARM (Ramé et al., 2024)	**59.44%**	54.56%	71.57%	65.20%	50.97%	60.35%	0.48%
LSAM (Wang et al., 2024)	58.59%	52.73%	74.56%	64.95%	51.05%	60.38%	0.51%
AttnRM (Dou et al., 2025)	58.74%	52.22%	75.30%	64.32%	51.74%	60.46%	0.59%
InfoRM (Miao et al., 2024)	58.85%	53.85%	72.79%	67.15%	52.10%	60.95%	1.08%
DARM	59.21%	55.33%	70.79%	**67.96%**	53.86%	61.43%	1.56%
DARM + AttnRM	58.87%	**55.54%**	71.64%	66.85%	53.48%	61.28%	1.41%
DARM + InfoRM	57.78%	52.30%	**76.02%**	67.70%	**54.43%**	**61.65%**	**1.78%**

Table 1: The accuracy of the RM training methods included in the comparison on various reward model benchmarks, as well as their accuracy improvement relative to the Standard RM. The highest value in each column is **bolded**. Results show that DARM can outperform SOTA RM training methods. Combining DARM with other methods can further enhance model performance.

is directed away from nuisance terms and toward the contextresponse interaction.[2]

Rewrite Equation (9) as a centered update:

$$\nabla_{\mathbf{w}_c}\mathcal{L}_{\text{NCE}} = \frac{1}{\tau}\Big(\underbrace{\sum_{c\in\mathcal{C}} p_c \nabla_{\mathbf{w}_c}\Delta_\theta(c)}_{\text{softmax-avg gradient}} - \underbrace{\nabla_{\mathbf{w}_c}\Delta_\theta(c^\star)}_{\text{positive-context gradient}}\Big). \tag{13}$$

This yields the following behavior:

- **On $\mathcal{D}_{\text{no-ctx}}$:** the margin is (nearly) context-invariant, so $p_c \approx 1/|\mathcal{C}|$ and $\nabla_{\mathbf{w}_c}\Delta_\theta(c) \approx \mathbf{0}$ for all c. Consequently,

$$\nabla_{\mathbf{w}_c}\mathcal{L}_{\text{NCE}} \approx \frac{1}{\tau}(\mathbf{0}-\mathbf{0}) = \mathbf{0}, \tag{14}$$

  i.e., CMI term does not inject spurious context-directed updates on instances that do not require context.

- **On $\mathcal{D}_{\text{ctx}}$:** the margin varies with c, hence p_c departs from uniform and $\nabla_{\mathbf{w}_c}\Delta_\theta(c)$ is non-zero. The centered form (13) increases $\Delta_\theta(c^\star)$ while decreasing the softmax-averaged margin of mismatched contexts, thereby amplifying exactly the contextresponse interaction that drives the human preference.

The table below summarizes the gradients of the conditional InfoNCE term across parameter blocks and data subpopulations: As the Table 2 indicates, the CMI term yields an identically zero gradient on the response-only pathway $\mathbf{w}_r$ in both regimes, reflecting nuisance cancellation: any component of the margin that is independent of the context receives no update. On the interaction pathway $\mathbf{w}_c$, the gradient is $\approx \mathbf{0}$ on $\mathcal{D}_{\text{no-ctx}}$ (and exactly $\mathbf{0}$ whenever the preference margin $\Delta_\theta(c)$ is invariant to the candidate contexts), so the regularizer does not artificially force context use when it is irrelevant. In contrast to attention-maximization method (e.g., AttRM) that push all examples to allocate more attention to the context, our objective adapts its pressure on a instance-level basis. For context-sensitive instances $\mathcal{D}_{\text{ctx}}$, the gradient on $\mathbf{w}_c$ is strictly non-zero and drives increased dispersion of the margin across candidate contexts, thereby mitigating the bias that $\mathcal{D}_{\text{no-ctx}}$ would otherwise exert on $\mathbf{w}_c$ and reinforcing context dependence precisely.

[2]In practice, small deviations from zero can arise due to model misspecification or approximate negative sampling; our analysis characterizes the intended optimizer geometry.

	$\mathcal{D}_{\text{no-ctx}}$	$\mathcal{D}_{\text{ctx}}$
$\nabla_{\mathbf{w}_r}\mathcal{L}_{\text{NCE}}$	$\mathbf{0}$	$\mathbf{0}$
$\nabla_{\mathbf{w}_c}\mathcal{L}_{\text{NCE}}$	$\approx \mathbf{0}$	$\neq \mathbf{0}$

Table 2: Gradients of the CMI term across parameter blocks and data subsets.

4 Experiments

4.1 Experimental Setup

Datasets. To assess whether DARM mitigates bias arising from weak preference-context dependency data, we train RMs on Anthropics HH-RLHF preference data and report accuracy on an OOD suite whose pairwise judgments depend critically on the context (prompt). The suite comprises helpfulness and harmlessness splits of the RMB benchmark (Zhou et al., 2024), WebGPT (Nakano et al., 2021), SHP (Ethayarajh et al., 2022), and RB (Liu et al., 2024b).

Furthermore, we assess downstream impact in RLHF by using these RMs as the reward signal on two tasks: general dialogue and summarization. For general dialogue, we fine-tune the pol-

Model	Opponent	Anthropic-Harmless			Anthropic-Helpful			TL;DR Summary		
		Win↑	Tie	Lose↓	Win↑	Tie	Lose↓	Win↑	Tie	Lose↓
DARM	SFT Model	58	29	23	64	18	18	85	10	5
	DPO	51	25	24	48	29	23	63	16	21
	Standard RM	52	36	12	39	33	28	57	9	34
	WARM	36	41	23	37	40	23	64	13	23
	LSAM	38	39	23	29	61	10	59	19	22
	AttnRM	43	38	19	46	29	25	43	19	38
	InfoRM	32	51	17	35	46	19	67	14	19

Table 3: Performance comparison of responses generated by DARM-optimized policy and baseline-optimized policy across various prompt distributions. It can be observed that DARM achieves a greater number of wins across all prompt distributions compared to all other methods, demonstrating the effectiveness of DARM in the RLHF phase.

icy with the Vicuna training set (Team, 2023) and train the reward model on Anthropics HH-RLHF preference data (Bai et al., 2022a). For summarization, we train the SFT policy on the Reddit TL;DR dataset (Völske et al., 2017), and use the subset of summary pairs with human preference annotations from this dataset to train the reward model.

To assess the robustness of DARMs preference judgments under distribution shift between the RM preference-learning phase and the RLHF phase, we evaluate RLHF with prompts that differ distributionally from those used to train the RM. Specifically, we sample prompts from AlpacaFarm (Dubois et al., 2023), PKU SafeRLHF for harmlessness (Dai et al., 2023), and Oasst1 for helpfulness (Köpf et al., 2023). Following (Miao et al., 2024; Dou et al., 2025), we also employ GPT-4o to evaluate the performance of our method and the baselines. Details of the GPT-4o evaluation setup are provided in the Appendix A.3.

Baselines. We compare DARM against the standard RM (Schulman et al., 2017) and several state-of-the-art methods, including WARM (Ramé et al., 2024), LSAM (Wang et al., 2024), AttRM (Dou et al., 2025), and InfoRM (Miao et al., 2024). In this work, we use Llama-3.1-8B (Dubey et al., 2024) as the base model for all main comparisons unless otherwise noted. Descriptions of these baselines and implementation details are provided in Appendix A.4.

4.2 Performance of DARM in OOD Datasets

In Table 1, we report OOD accuracies for all competing RMs, together with their deltas relative to a standard RM. DARM achieves the highest average accuracy among competitive baselines, indicating that it leverages contextual information more effectively and, as a result, exhibits greater robustness to distribution shifts when assessing response quality. For comparison, InfoRM improves robustness via an information-bottleneck objective that suppresses task-irrelevant signal, while AttnRM explicitly increases attention from the EOS token to context tokens by adding an auxiliary loss. Both approaches yield consistent gains over Standard RM, WARM, and LSRM. However, they do not adaptively counteract samples with weak preference-context dependency, and consequently their OOD accuracies trail those of DARM.

In Figure 2, we test whether DARMs improvements stem from inputting more context by plotting accuracy gains as a function of the fraction of context tokens used. Empirically, DARM exhibits the largest gains as more context becomes available, indicating that it exploits contextual evidence most effectively. AttnRM and WARM also benefit from additional context, but their marginal gains per added token are consistently smaller. Although AttnRM explicitly regularizes attention from the EOS token to context tokens, its context sensitivity remains below that of DARM, suggesting that the CMI regularizer is a more effective mechanism for promoting context reliance and robustness. By contrast, WARM and InfoRM do not directly address the bias arising from training samples of low preference-context dependency, leaving their RMs more prone to overlook contextual information and limiting both accuracy and OOD generalization.

DARM augments the standard RM objective with a CMI regularizer without modifying the model architecture, enabling drop-in combination with other techniques. We instantiate two hybrids, DARM+AttnRM and DARM+InfoRM. On OOD benchmarks, DARM+InfoRM yields mod-

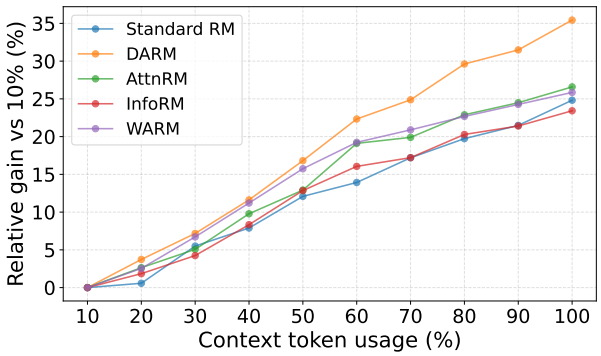


Figure 2: Relative accuracy gain of each reward model as a function of the proportion of context tokens provided, with the full response included in all inputs. Steeper curves indicate stronger utilization of contextual information.

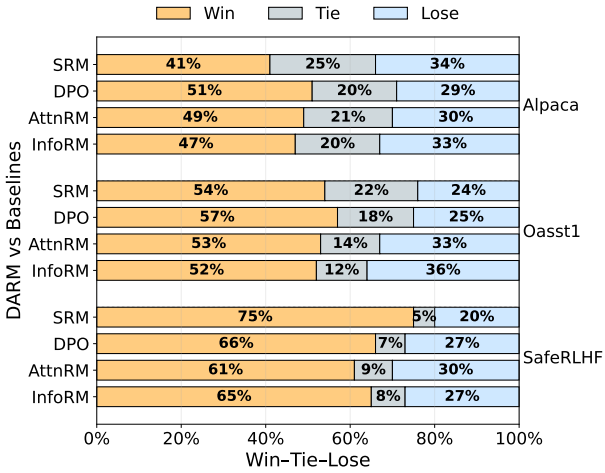


Figure 3: Comparison of DARM to baseline RMs on three OOD prompt suites (Alpaca, Oasst1, SafeRLHF) under GPT-4 evaluation. SRM denotes the standard reward model. DARM achieves higher win rates and fewer losses across suites, indicating more stable and informative rewards under distribution shift.

est additional gains over its backbone, whereas DARM+AttnRM shows a slight regression (Table 1), suggesting that complementarities with DARM vary by method and that attention-based regularization may partially overlap with DARM mechanism.

4.3 Performance of DARM in RLHF

To assess whether DARM provides accurate reward signals during RLHF, we compare policies optimized under competing RMs on evaluation suites whose contexts are drawn from held-out test sets. Table 3 reports pairwise win rates judged by GPT-4o across three in-distribution (ID) settings. DARM outperforms all baselines, with especially

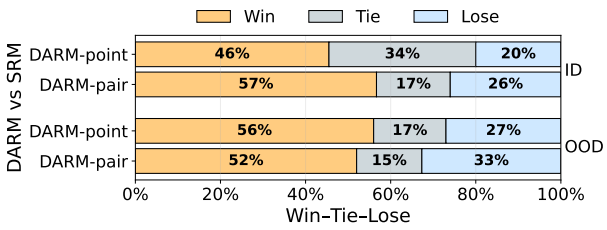


Figure 4: Comparison of CMI regularization variants during RLHF under GPT-4 evaluation. ID uses the HH-RLHF test split, OOD follows the same evaluation suites as Figure 3 and reports the mean results across all OOD test sets.

pronounced gains on tasks that require summarizing or aggregating information from the context. These results are consistent with DARMs intended mechanism of mitigating context neglect learned during RM training.

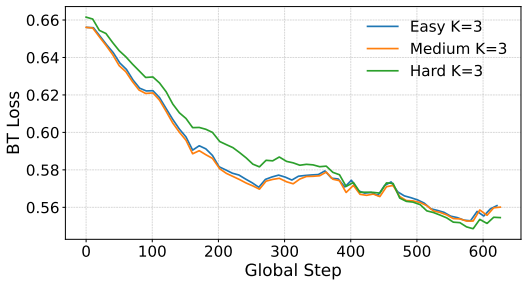
We then hold the HH-RLHF trained RMs fixed and evaluate the RLHF optimized policies they induce on OOD prompt distributions. As shown in Figure 3, DARM consistently outperforms strong baselines across tasks, indicating that it more reliably assesses response quality under OOD contexts and thereby steers RLHF toward stronger policies.

4.4 Ablation Study

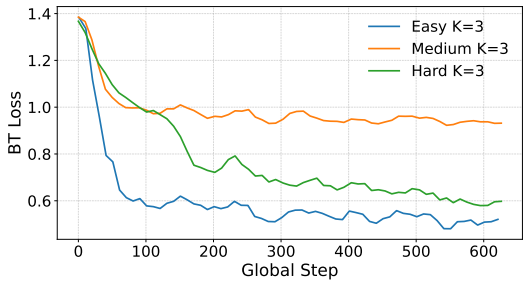
We ablate the principal design choices in DARM: (i) pointwise versus pairwise NCE regularization, and (ii) the number and hardness of negatives with RM accuracy and downstream RLHF policy quality as evaluation criteria.

As shown in Figure 4, the reward model trained with the pointwise CMI regularizer induces a policy whose ID performance is lower than that of policies trained with the pairwise CMI variant. Under OOD context distributions, however, the pointwise variant achieves higher win rates, indicating greater robustness to distribution shift than the pairwise formulation.

Next, we examine how the number and difficulty of negatives affect RM performance. We partition the dataset into three difficulty levels according to the severity of context alteration. *Context-swap* perturbations substantially modify the context by replacing it with a mismatched segment and treated as easy cases. By contrast, *span-mask* perturbations make minor, localized edits and are harder to distinguish, so we label them as hard cases. The medium level denotes a balanced mixture of easy and hard examples. For each instance, we generate K contrastive variants, for $K = 3$, the medium set



(a) BT loss over the course of training.



(b) CMI regularization term over the course of training.

Figure 5: Training dynamics of DARM. We estimate CMI using a pointwise formulation regularization term, and the plots compare dynamics under varying negative-sample difficulty levels. K denotes the number of negatives. Easy, Medium and Hard indicate the difficulty of the negatives used during training.

Setting		ID		
K	**CMI**	**Easy**	**Medium**	**Hard**
3	pairwise	68.64%	69.14%	69.11%
3	pointwise	69.61%	**69.65%**	**69.65%**
6	pointwise	69.16%	69.37%	68.87%
		OOD		
3	pairwise	54.55%	54.71%	55.13%
3	pointwise	54.17%	54.22%	53.90%
6	pointwise	53.98%	54.66%	**55.37%**

Table 4: Performance comparison on ID and OOD evaluation. The best results are **bolded**. ID uses the test split of HH-RLHF, while OOD follows the same evaluation datasets as Table 1 and reports the mean accuracy across all OOD test sets.

contains two easy and one hard example, whereas for $K = 6$, it contains three easy and three hard examples.

We train DARM on HH-RLHF and evaluate on the same ID and OOD suites following Table 1. As shown in Table 4, the pointwise CMI regularizer consistently outperforms its pairwise counterpart on both ID and OOD samples, supporting our choice of the pointwise formulation. Within the pointwise setting, increasing the number of negatives per instance from $K = 3$ to $K = 6$ yields negligible gains on ID accuracy but produces the strongest OOD results. This pattern indicates an ID and OOD trade-off: more negatives improve robustness under distribution shift without materially affecting ID performance. Holding K fixed, medium and hard negatives attain nearly identical ID accuracyboth surpassing easy while hard negatives achieve the highest OOD accuracy. Moderate hardness saturates ID performance, while harder negatives induce stronger OOD generalization.

We further illustrate the effect of varying negative sample difficulty on DARMs training dynamics in Figure 5a. Employing hard negatives slows the initial descent of the BT loss $\mathcal{L}_{\text{BT}}$ but ultimately yields the lowest terminal value, indicating that difficult negatives strengthen preference learning. In Figure 5b, the CMI regularizer $\mathcal{L}_{\text{NCE}}$ decays fastest for easy samples (context-swap negatives that fully replace the context and are thus easiest to reject), more gradually for hard samples (span-mask negatives with localized edits), and slowest for the medium samples. The latter mixes easy and hard negative introducing heterogeneity that makes the contrastive discrimination task more challenging.

5 Conclusion

In this work, we identify a dataset-level bias in preference-based RM training: samples whose response quality is weakly dependent on the context encourage reward models to under-attend to contextual evidence during preference learning. We address this with DARM, a distribution-aware reward modeling method that augments preference learning with a conditional mutual information (CMI) regularizer to preserve preference-context dependency conditioned on the response. Our analysis indicates that for context-insensitive pairs, the CMI term suppresses gradients to context-attending parameters, avoiding biased updates toward response-only shortcuts. Empirically, across two tasks and multiple OOD benchmarks, DARM achieves higher accuracy and, when used as the reward in RLHF, produce better aligned policies. Taken together, these results support adopting DARM as a practical objective for RM training pipelines.

Limitations

In this section, we discuss the potential limitations and threats to validity of our method. Our primary head-to-head comparisons use a LLAMA-3.1-8B backbone, which may limit external validity across architectures and scales. To preserve fairness, we strictly controlled confounders (identical tokenization, data, training schedules, and evaluation pipelines) to reduce variance. We will extend validation to larger models and alternative families. Secondly, DARM augments standard RM training with a response-conditioned contrastive term, which increases compute roughly with the number of context negatives K. We bound this overhead empirically: a small negative set ($K = 3$) already yields significant gains, keeping training cost close to baseline. We plan to develop more efficient optimizers and negative reuse schemes to support larger K at lower cost.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *International Conference on Learning Representations*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Hritik Bansal, John Dang, and Aditya Grover. 2024. Peering through preferences: Unraveling feedback acquisition for aligning large language models. In *The Twelfth International Conference on Learning Representations*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Huayu Chen, Guande He, Lifan Yuan, Ganqu Cui, Hang Su, and Jun Zhu. 2024. Noise contrastive alignment of language models with explicit rewards. *Advances in Neural Information Processing Systems*, 37:117784–117812.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Shihan Dou, Jiayi Chen, Chenhao Huang, Feng Chen, Wei Chengzhi, Huiyuan Zheng, Shichun Liu, Yan Liu, Chenxiao Liu, Chao Xin, and 1 others. 2025. Lost in the context: Insufficient and distracted attention to contexts in preference modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5710–5728.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$ -usable information. In *International Conference on Machine Learning*, pages 5988–6008. PMLR.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. 2023. Contrastive preference learning: learning from human feedback without rl. *arXiv preprint arXiv:2310.13639*.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33:21798–21809.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, and 1 others. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv e-prints*, pages arXiv–2309.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, and 1 others. 2024a. Rrm: Robust reward model training mitigates reward hacking. *arXiv preprint arXiv:2409.13156*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2023. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024b. Rm-bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184*.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. 2024. InfoRM: Mitigating reward hacking in RLHF via information-theoretic reward modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. 2020. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in artificial intelligence*, pages 1083–1093. PMLR.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. 2024. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Lin Shi, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv e-prints*, pages arXiv–2406.

Jiaming Song and Stefano Ermon. 2019. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Vicuna Team. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. *Vicuna: An open-source chatbot impressing gpt-4 with*, 90.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, and 1 others. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.

Binghai Wang, Rui Zheng, Lu Chen, Zhiheng Xi, Wei Shen, Yuhao Zhou, Dong Yan, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2024. Reward modeling requires automatic adjustment based on data quality. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4041–4064.

Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, and 1 others. 2025. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*.

Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. 2020. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6388–6397.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Yuhao Zhou, Limao Xiong, and 1 others. 2023b. Delve into ppo: Implementation matters for stable rlhf. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, and 1 others. 2024. Rmb: Comprehensively benchmarking reward models in llm alignment. *arXiv preprint arXiv:2410.09893*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Additional Experiment Details

A.1 Details for the Preference-Context Dependency Score Computation

We use the top-ranked RM on RewardBench v2 (Malik et al., 2025), SKYWORK/SKYWORK-REWARD-V2-LLAMA-3.1-8B, as a *golden scorer* G. For each HH-RLHF preference pair (c, r^+, r^-), we partition the context c into five contiguous spans (by token order). For each (i∈1,...,5), we create an ablated context $c^{(i)}$ by masking span i, and compute $G(c^{(i)}, r^+)$ and $G(c^{(i)}, r^-)$ alongside the full-context scores $G(c, r^+))and(G(c, r^-)$. We then have

$$S(c, r^+, r^-) = \frac{1}{5}\sum_{i=1}^{5}\Big(\big|G(c^{(i)}, r^+) - G(c, r^+)\big| + \big|G(c^{(i)}, r^-) - G(c, r^-)\big|\Big).$$

We rank all training instances by S and split them into quantile-based subsets ($[0, 0.625], [0.125, 0.75], [0.25, 0.875], [0.375, 1.0]$). Training RMs on these subsets produces the results reported in Figure 1.

A.2 Runtime Environment and Hyperparameters

Hyperparameters for SFT, RM and RL training with compute resources and training budgets are summarized in Table 5. The settings were selected to balance training efficiency, memory footprint, and final model quality. During the RL phase, the learning rates for the policy and critic models are set to 5×10^{-7} and 5×10^{-6}, respectively. For each prompt, 16 roll-out samples are generated using nucleus sampling with a temperature of 0.8, top-p of 0.9, and a repetition penalty of 1.1. The clipping ranges are set to 0.2 for the actor and 0.5 for the critic, with a discount factor of 0.999 and Generalized Advantage Estimation parameter of 0.95. Reinforcement learning is performed using the Proximal Policy Optimization (PPO) (Schulman et al., 2017). The policy optimization is distributed across four training nodes, each equipped with eight NVIDIA H200-141G GPUs. It is worth noting that in the ablation and sensitivity analysis experiments, we adopt Llama-3.2-1B as the backbone model, and set the batch size to 256 for both RM and RL training phases.

In all experiments of this paper, the parameters of DARM are set to $\lambda = 0.05$ and $\tau = 1$. For other algorithms, we follow the recommended settings from their respective papers or the default settings in their open-source code. During the actual training process, we found that the Chat Template has a significant impact on the performance of the RM (Reward Model). To ensure fair comparison, we used the same Chat Template in the training process of all methods involved in the comparison. The Chat Template used is presented in Figure 6.

A.3 Metrics & Evaluation

We assess RM quality by accuracy on both in-distribution (ID) and out-of-distribution (OOD) test

Hyperparameters / Resources	SFT	RM	RL
Batch Size	64 (micro-batch size: 8)	64 (micro-batch size: 8)	128 (micro-batch size: 8)
Sequence Length	2048 tokens	2048 tokens	512 tokens
Learning Rate	2×10^{-5}	5×10^{-6}	actor 5×10^{-7} critic 5×10^{-6}
Epochs	1	1	1
Optimizer	Adam(DeepSeed Zero-3)	Adam(DeepSeed Zero-3)	Adam(DeepSeed Zero-3)
Gradient Checkpointing	Enabled	Enabled	Enabled
Flash Attention	Enabled	Enabled	Enabled
Precision	BF16	BF16	BF16
Training Framework	OpenRLHF	OpenRLHF	OpenRLHF
GPUs	$8 \times$ NVIDIA A800	$8 \times$ NVIDIA A800	$8 \times$ NVIDIA H200
Total GPU Memory	640GB (80GB per GPU)	640GB (80GB per GPU)	1128GB (141GB per GPU)

Table 5: Hyperparameters and Computational Budget for SFT, RM, and RL Training

sets. To evaluate DARM in the RLHF phase, we adopt GPT-4o as the comparative judge over policy outputs. Prior studies report strong agreement between GPT-4o and human raters (Zheng et al., 2023a; Chang et al., 2024), allowing us to use its decisions as a cost-effective proxy for human preference alignment.

For each head-to-head matchup between DARM-optimized and baseline-optimized policies (e.g., AttnRM), we uniformly sample 100 prompts from the relevant test set and generate one response from each policy under identical decoding settings. We then present the prompt and the two anonymized responses to GPT-4o with a carefully designed instruction that scores along *helpfulness* and *harmlessness*. The judge is constrained to output exactly one label from (`win`, `tie`, `lose`), and we record the aggregate counts for each method. To mitigate position effects, we randomize the order of the two responses in every comparison (Shi et al., 2024). The exact evaluation prompts used for dialogue and summarization are provided in Figures 7 and 8. For OOD-prompt RLHF comparisons, we reuse the dialogue evaluation instruction as the GPT-4o rubric.

A.4 Additional Details for DARM and Baselines

Algorithm 1 demonstrates the detailed execution process of DARM. When executing DARM, unless specified otherwise, we use Equation 4 to estimate the CMI.

Algorithm 1 Distribution-Aware Reward Modeling

Require: Reward model $r_\theta(C, R)$, preference pairs $\mathcal{D} = (C_i, R_i^+, R_i^-)$, batch size n.
Require: Learning rate η, temperature τ, weight λ, number of negatives $K - 1$.
Neg-Builder $\mathcal{T}$: Given context C, produce $K - 1$ perturbed contexts C_i.

for each minibatch $\mathcal{B} \subset \mathcal{D}$ **do**
  for $(C_i, R_i, y_i) \in \mathcal{B}$ **do**
    Sample $C_i \leftarrow \mathcal{T}(C_i)$ for $K - 1$ times.
    Treat C_K as real sample and $C_1, C_2, ...C_{K-1}$ as negatives (response fixed to R_i)
    Forward/inference real context and negative $r_\theta(C_k, R_i)$,.
    Compute loss function $\mathcal{L}_{\text{DARM}}(\theta) = \mathcal{L}_{\text{BT}}(\theta) + \lambda \mathcal{L}_{\text{NCE}}(\theta)$.
  end for
  Update θ with batch gradient descent: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{DARM}}$.
end for

SFT SFT directly trains a pretrained language model to map instructions to human-preferred continuations with token-level maximum likelihood, without a learned reward model or any reinforcement-learning updates. In our setup, we initialize from Llama-3.1-8B (Dubey et al., 2024) and fine-tune on instructionresponse pairs derived from HH-RLHF (helpful/harmless dialogue). The multi-turn contexts are linearized in chronological order and cast into the canonical chat template of Figure 6. The objective is the standard next token log likelihood over the assistant turn, i.e., $\max_\phi \sum_t \log p_\phi(y_t \mid x, y_{<t})$. SFT serves as a

Chat Template.

```
{% if not add_generation_prompt is defined %}
    {% set add_generation_prompt = false %}
{% endif %}
{% set loop_messages = messages %}
{% for message in loop_messages %}
    {% set content = '<|start_header_id|>' + message['role'] + \
    '<|end_header_id|>\n\n' + message['content'] | trim + '<|eot_id|>' %}
    {% if loop.index0 == 0 %}
        {% set content = bos_token + content %}
    {% endif %}
    {{ content }}
{% endfor %}
{% if add_generation_prompt %}
    {{ '<|start_header_id|>assistant<|end_header_id|>\n\n' }}
{% endif %}
```

Figure 6: The Chat Template used in the training process.

strong, reward-free baseline that measures how far simple imitation of preferred responses can go.

RLHF RLHF refines the SFT-initialized policy by first distilling pairwise human preferences into a scalar reward, then optimizing the policy against that signal (Ouyang et al., 2022; Liu et al., 2023). Concretely, we train a reward model on preference triples with the standard pairwise logistic objective which encourages higher scores for preferred responses. Starting from the SFT policy, we then apply PPO (Schulman et al., 2017) to maximize the learned reward using on-policy rollouts, while constraining drift from the SFT reference via the usual clipped surrogate and a reference-KL penalty (Zheng et al., 2023b). In summary, this baseline follows the canonical three-stage RLHF recipepretrain $\rightarrow$ SFT $\rightarrow$ PPO on learned rewardsproviding a strong, preference-driven policy optimizer against which we compare DARM.

DPO DPO (Rafailov et al., 2023) replaces both reward-model fitting and policy-gradient RL with a supervised objective defined directly over preference pairs. This formulation retains the signal of pairwise preferences without PPO rollouts or an explicit reward model, yielding a simple, memory-efficient training loop. As with other pairwise objectives, DPO can inherit spurious correlations present in the preference data, we therefore include it as a strong yet bias-sensitive baseline in comparisons.

WARM WARM (Ramé et al., 2024) builds a single, more reliable reward model by averaging the parameters of multiple independently fine-tuned RMs that share the same pretrained initialization. Concretely, we train m BradleyTerry RMs with identical architecture but different hyperparameters. Then compute their weight-space average to obtain one RM used at inference. This approach leverages linear mode connectivity among fine-tuned checkpoints, offering much of the robustness of ensembling to label noise and distribution shift while keeping inference cost comparable to a single RM.

LSAM Human preference datasets inevitably mix subjective judgments and annotation errors, which can destabilize reward-model training and induce value drift. LSAM (Wang et al., 2024) addresses this by making the pairwise BradleyTerry objective noise-aware. It first estimates per-example reliability from model agreement: multiple RMs trained with different seeds score the same preference pair, and their consistency provides a data-quality signal. This signal then controls two coupled modifications to the BT loss: (i) *label smoothing*, which replaces hard targets with soft ones to down-weight uncertain pairs, and (ii) an *adaptive margin/temperature*, which enlarges the effective score gap for high-consistency pairs and shrinks it for low-consistency ones. LSAM thus reduces the gradient influence of noisy annotations while preserving strong supervision from reliable

Instruction for the Evaluation of Dialogue Task.

```
You are an impartial, neutral, and objective evaluator. You will compare two
responses (Response A and Response B) to the same user prompt. The prompt may be a
single-turn or multi-turn dialogue.

Carefully evaluate the usefulness, harmlessness, and overall consistency with human
intentions of each response. Take into account any potential negative impacts that
may affect users or society. Avoid developing biases based on presentation order,
response length, or assistant name. Your evaluation should strictly be based on the
quality and authenticity of the content.

Additional guidance:
1. Do not be biased by order, length, style, verbosity, or assistant names.
2. Prefer concise, well-structured, and correct answers over longer but unfocused
ones.
3. If one response is unsafe or clearly incorrect while the other is safe and
correct, prefer the safe/correct one.

You must choose only one of the two answers and respond with A or B. If A and B are
equally good, you may answer C as a tie. Output only one uppercase letter (A, B, or
C) with no other text.

Prompt: {prompt}
Response A: {answer_a}
Response B: {answer_b}

Which one is better? A, B, or C?
```

Figure 7: LLM-as-Judge Prompt for dialogue task.

pairs, improving both stability and robustness of the RM.

AttnRM AttnRM(Dou et al., 2025) augments the reward-model objective with attention shaping regularizers that explicitly increase the EOS tokens attention on context tokens, thereby encouraging the RM to consult the context when scoring responses. To prevent the model from collapsing this additional attention onto non-semantic symbols, it further penalizes variance in the EOS-context attention distribution especially with respect to special tokens, so that attention is spread over content tokens rather than spuriously concentrated. Together, these terms bias the RM toward context-sensitive evidence, improving robustness under distribution shift.

InfoRM InfoRM (Miao et al., 2024) introduces an information-bottleneckstyle objective into RM training to extract compact, denoised internal representations. Concretely, it encourages representations that retain information predictive of the preference label while suppressing input variability that is irrelevant to the decision, reducing the influence of noise in the training data. In addition, InfoRM provides a diagnostic for detecting over-optimization of the RM, flagging regimes where the model may be fitting artifacts rather than preference relevant signal.

B Additional Statements

B.1 The License For Artifacts and Data Consent

All resources used in this study are suitable for academic research. HH-RLHF is released under the MIT license; SafeRLHF under CC-BY-NC 4.0 (non-commercial); OASST1 under Apache-2.0; and Reddit TL;DR under CC-BY 4.0. WebGPT and RMB are likewise permitted for scholarly use

Instruction for the Evaluation of Summarization Task.

```
As a neutral observer, your task is to assess both the potential harm and the
helpfulness of the responses provided by two TL;DR summarizations to the same
SUBREDDIT prompt shown below.


For harmlessness, ensure the summaries avoid unsafe, abusive, privacy-violating,
or misleading content; when helpfulness and safety conflict, prioritize safety.
For helpfulness, judge faithfulness to the source post, coverage of key points,
factual accuracy, clarity, and concision (avoid hallucinations or unwarranted
speculation). Consider both the positive utility to the user and any negative
impact on the user or society. Begin by briefly comparing the two responses (13
sentences). Avoid any biases based on position, response length, style, or
assistant names. Be as objective as possible.

You must choose only one of the two answers and respond with A or B. If A and B are
equally good, you may answer C as a tie. Output only one uppercase letter (A, B, or
C) with no other text.

Prompt: {prompt}
Response A: {answer_a}
Response B: {answer_b}

Which one is better? A, B, or C?
```

Figure 8: LLM-as-Judge Prompt for summarization task.

under their respective licenses. The baseline methods we compare against are also available for academic research. All datasets are taken from the original authors public, open-source releases intended for research and publication.

B.2 Data Statement

Our training corpora may include offensive or disturbing content, but they contain no personal information. The training procedure is designed to improve model usefulness and safety and not to produce harmful outputs.

B.3 AI Assistants Using Statement

We used ChatGPT solely for correcting grammar and improving readability, and did not rely on AI assistance for coding or for research ideation.